Individual responses made easy

**Will G. Hopkins**
Victoria University, Melbourne, Australia

WE ALL INHERIT AND ACQUIRE different characteristics. When we experience a treatment aimed at changing our physiology, these characteristics may modify the effect of the treatment, making it more or less beneficial, harmful, or ineffective in different individuals. The issue of individual responses to treatments is therefore one of the most important in experimental research, yet few researchers acknowledge the issue in their published studies, and attempts to quantify individual responses are rare and usually deficient. This climate of neglect and ignorance needs to change, especially now that genome sequencing and pervasive monitoring of individuals can provide researchers with the subject characteristics that account for individual responses and allow more efficient ethical targeting of treatments to individuals.

The synthesis review by Hecksteden and colleagues (2) in this issue of the *Journal of Applied Physiology* is timely, because it deals with some of the methodological challenges in quantifying individual responses in parallel-group controlled trials, where an experimental and control group are measured before and after their respective treatments. The authors assert that proper quantification of individual responses requires repeated administration of the experimental treatment to determine the extent to which each individual’s response consists of reproducible and random components additional to the random variation due to error of measurement experienced equally by all individuals. In essence, the reproducible individual responses are those that could be explained by differences between subjects in inherited and acquired stable characteristics or traits, whereas the apparently random responses could be due to changes in subject characteristics or states between administrations of the treatment. Although two or more administrations are indeed required to partition individual responses into these two components, partitioning is possible only if effects of the treatment wash out fully between administrations. Repeated administrations therefore make sense for acute effects of short-term treatments, where the treatments should be administered in crossover fashion (7). For training and other long-term treatments, repeated administrations are seldom logically feasible, nor are they even possible in principle, if the intent of the study is to provide evidence for a permanent change in behavior or physiology. In any case, for long-term treatments, subject states average out to become subject traits, so one can expect the random component of individual responses due to subject states to become trivial or meaningless.

The Hecksteden et al. synthesis review is based on sound but challenging statistical principles, and the focus on repeated administration of treatments may distract researchers from the straightforward and legitimate analysis of individual responses to a single administration of a treatment. The authors have referenced several of my own publications on the appropriate methods (3-6), but I will update and summarize the methods here, in the hope of improving the analysis and reporting of controlled trials at least in this journal.

Individual responses are manifest as a larger standard deviation of the change scores in the experimental group than in the control group. It is therefore imperative that researchers report the standard deviations of the change scores, along with the means. This simple requirement will also provide all the inferential information needed for inclusion of the study in meta-analyses not only of the mean effect of the treatment but also of the individual responses.

The individual responses are summarized by a standard deviation (SDIR) given by the square root of the difference between the squares of the standard deviations of the change scores in the experimental (SDExp) and control (SDCon) groups: 

$$SDIR = \sqrt{(SDExp^2 - SDCon^2)}$$

One should consider this standard deviation to be the amount by which the net mean effect of the treatment differs typically between individuals. Confidence limits for the standard deviation are obtained by assuming its sampling variance is normally distributed, with standard error given by 

$$\sqrt{[2(SEExp^2/DFExp + SDExp^2/DFCon)]}$$

where DFExp and DFCOn are the degrees of freedom of the standard deviations in the two groups (usually their sample sizes minus 1). The upper and lower confidence limits for the true value of the variance of individual responses (SDIR^2) are given by its observed value plus or minus this standard error times 1.65, 1.96, or 2.58 for 90, 95, or 99% confidence limits, respectively (5). This formula is the basis for the confidence limits provided via mixed modeling with a procedure such as Proc Mixed in the Statistical Analysis System (SAS Institute, Cary, NC). Negative values of variance for the observed individual responses and the confidence limits can occur, especially when there is large uncertainty arising from a small sample size or a large error of measurement. Taking the square root of a negative number is not possible, so I advocate changing the sign first, then presenting the result as a negative standard deviation, interpreted as more variation in the control group than in the experimental group. If the upper confidence limit is negative, the researcher should consider explanations beyond mere sampling variation, such as compression of the responses in the experimental group arising from the treatment bringing subjects to a similar level.

The magnitude of the standard deviation and its confidence limits should also be interpreted. The default approach for interpretation of the change in a mean is standardization, where the mean change is divided by the standard deviation of all subjects at baseline (before the treatment). The thresholds for interpreting the standardized mean change (0.2, 0.6, 1.2, 2.0, and 4.0 for small, moderate, large, very large, and extremely large) (6) need to be halved (0.1, 0.3, 0.6, 1.0, and 2.0) for interpreting the magnitude of effects represented by standardized standard deviations (8), including individual responses. Halving the thresholds implies that the magnitude of the standard deviation is evaluated as the difference between a
typically low value (mean \(\pm\) SD) and a typically high value (mean + SD).

To exemplify this approach, I will draw on data for effects of training on maximum oxygen uptake (\(V_{\text{O}2\text{max}}\)) in one of the classic studies of Bouchard and colleagues (1) cited by Hecksteden et al. They reported a change of 384 ± 202 ml/min (mean ± SD) in an experimental group of 720 sedentary young adults. Means and SD at baseline were not shown, although the median was 2,159 ml/min. Let us assume a baseline of 2,200 ± 330 ml/min (mean ± SD). There were no data for changes in a control group, but the “coefficient of variation” for \(V_{\text{O}2\text{max}}\) was stated to be 5%, presumably based on two tests separated by \(\sim\)1 wk. The standard deviation of change scores over this short period would therefore be \(5\sqrt{2}\) or 7.1% and probably somewhat larger over the 20 wk of the training study. Let us assume 8%, or 176 ml/min. Evidently much of what seems like individual responses was nothing more than noise, and the SD of individual responses to training free of this noise was only \(\sqrt{(202^2 - 176^2)} = 99\) ml/min. The lower and upper 90% confidence limits, assuming 50 subjects contributed to the estimate of the coefficient of variation for \(V_{\text{O}2\text{max}}\), were \(-33\) and 144 ml/min. In standardized units (dividing the values by the assumed baseline SD of 330 ml/min), the individual responses were 0.30 (90% confidence limits \(-0.10\) and 0.44). Thus the observed individual responses were borderline small, but given the uncertainty, the individual responses could have been trivial or at most moderate. Assuming a control group showed no mean change over 20 wk, the standardized mean effect itself was 384/330 = 1.16 or borderline moderate-large. The overall effect of training, after removing the effects of noise, could be summarized as 384 ± 99 (mean ± SD for individual responses) or 1.16 ± 0.30 in standardized units. Thus ignoring the uncertainties in the mean and SD in this sample, the effect on individuals ranged typically from moderate (mean \(\pm\) SD = 0.86) to large (mean + SD = 1.46).

The uncertainty in the estimate of individual responses in the above example was large, disappointingly so in view of the assumed reasonably large sample size in the control (50) and the unusually large sample size in the training group (720, which is effectively infinite compared with 50). The main source of the uncertainty is the error of measurement (5%), which is somewhat larger than the smallest important effect (0.2 of the assumed 330 ml/min, or 3%). Evidently investigation of individual responses will be successful only with good sample sizes and measures with low noise. As suggested by Hecksteden et al., averaging repeated measurements on each subject before and after the treatments is one solution to the problem of low sample size and noisy measures.

Hecksteden et al. also reviewed the approaches to identifying positive responders, nonresponders, and negative responders. The uncertainty in each change-score response can be expressed as confidence limits by multiplying the appropriate value of the \(t\) statistic by SD_{crew} (or by \(\sqrt{2}\) times the standard error of measurement from a relevant reliability study) (3). For many outcome measures, the uncertainty is considerably greater than the smallest important change, so even an inactive control treatment will appear to produce positive and negative responses. True proportions of positive and negative responses are therefore provided by counts of observed responses only rarely, when the error of measurement is much less than the smallest important change. The proportions can be estimated otherwise by assuming the responses have a \(t\) distribution defined by the net mean response and the standard deviation for individual responses, and confidence limits for the proportions could be derived by bootstrapping. I do not advise such calculations when the lower confidence limit for the standard deviation is negative.

Finally, Hecksteden et al. provide advice on the complexities of analytical models that include subject characteristics to account for individual responses. In my view these analyses need not be so daunting: it is enough to include each subject characteristic separately as a numeric linear or nominal effect interacted with the group effect to predict the change scores in each group. The predictor can represent a trait or a state of the subjects, in which case it is a moderator of the treatment effect. The predictor can also be a change score for a physiological state variable, in which case its effect represents potential mediation of the treatment effect and/or its individual responses (best understood with a scatterplot of change scores of dependent vs predictor). To the extent that the predictor has a substantial effect, the standard deviation of individual responses will decrease substantially. My spreadsheet for analysis of controlled trials allows for such analyses and provides confidence limits and appropriate evaluation of all effects (5). When subject characteristics are correlated and the researcher is interested in the unique contribution of each, a statistical package is needed to perform multiple linear regression, with each characteristic included as a main effect. Models that include interactions of subject characteristics should be considered only if there is a good theoretical basis for such interactions.

In summary, adequate quantification of individual responses in a controlled trial requires a large sample size or averaging of repeated measurements to compensate for a large error of measurement. The write-up should include means and standard deviations of change scores in all treatment groups. Individual responses should be reported as a standard deviation with confidence limit derived from the standard deviations of the change scores before and after inclusion in the analysis of any subject characteristics representing potential moderators and of any change scores representing potential mediators. The standard deviation at baseline should be used to assess the magnitudes of effects and individual responses by standardization. These recommendations will be novel for most researchers, but they are not especially rocket science, and they need to be implemented.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: W.G.H. conception and design of research; W.G.H. analyzed data; W.G.H. drafted manuscript; W.G.H. edited and revised manuscript; W.G.H. approved final version of manuscript.

REFERENCES


